Web Crawling

Introduction to Information Retrieval
INF 141/ CS 121
Donald J. Patterson

Content adapted from Hinrich Schütze
http://www.informationretrieval.org
Overview

- Introduction
- URL Frontier
- Robust Crawling
  - DNS
Introduction

The User
flickr:crankyT

Search Results

Search
Go Search

Web Spider

The Web

Indexer

Indices

Ad Indices

Indices
The basic crawl algorithm

- Initialize a queue of URLs ("seed" URLs)
- Repeat
  - Remove a URL from the queue
  - Fetch associated page
  - Parse and analyze page
  - Store representation of page
  - Extract URLs from page and add to queue
Crawling the web

Introduction

Seed Pages

Web Spider

Crawled Pages

URL Frontier

The Rest of the Web
Basic Algorithm is not reality...

- Real web crawling requires multiple machines
- All steps distributed on different computers
- Even Non-Adversarial pages pose problems
  - Latency and bandwidth to remote servers vary
  - Webmasters have opinions about crawling their turf
    - How “deep” in a URL should you go?
- Site mirrors and duplicate pages
- Politeness
  - Don’t hit a server too often
Basic Algorithm is not reality...

- Adversarial Web Pages
- Spam Pages
- Spider Traps (http://www.devin.com/sugarplum/)
  - Find crawlers that don’t respect robots.txt files
    - spammers
- Griefing
- Naturally occurring
  - Calendars
Minimum Characteristics for a Web Crawler

- Be Polite:
  - Respect implicit and explicit terms on website
  - Crawl pages you’re allowed to
  - Respect “robots.txt” (more on this coming up)

- Be Robust
  - Handle traps and spam gracefully

Introduction
Desired Characteristics for a Web Crawler

• Be a distributed systems
  • Run on multiple machines
• Be scalable
  • Adding more machines allows you to crawl faster
• Be Efficient
  • Fully utilize available processing and bandwidth
• Focus on “Quality” Pages
  • Crawl good information first
Desired Characteristics for a Web Crawler

- Support Continuous Operation
- Fetch fresh copies of previously crawled pages
- Be Extensible
  - Be able to adapt to new data formats, protocols, etc.
  - Today it’s AJAX, tomorrow it’s HTML6, then....
Updated Crawling picture

URL Frontier

Crawled Pages

Seed Pages

URL Frontier "Priority Queue"

The Rest of the Web

Spider Thread
• Frontier Queue might have multiple pages from the same host
• These need to be load balanced (“politeness”)
• All crawl threads should be kept busy
Politeness?

- It is easy enough for a website to block a crawler
- Explicit Politeness
  - “Robots Exclusion Standard”
    - Defined by a “robots.txt” file maintained by a webmaster
    - What portions of the site can be crawled.
      - Irrelevant, private or other data excluded.
    - Voluntary compliance by crawlers.
    - Based on regular expression matching
Politeness?

- Explicit Politeness
- “Sitemaps”
  - Introduced by Google, but open standard
  - XML based
  - Allows webmasters to give hints to web crawlers:
    - Location of pages (URL islands)
    - Relative importance of pages
    - Update frequency of pages
  - Sitemap location listed in robots.txt
Politeness?

- Implicit Politeness
  - Even without specification avoid hitting any site too often
  - It costs bandwidth and computing resources for host.
## URL Frontier

### Politeness?

**Statistics for:**
djp3.net

**Summary**
- **When:** Monthly history, Days of month, Days of week, Hours
- **Who:** Countries, Full list, Last visit, Unresolved IP Address
- **Robots/Spiders visitors:** Full list, Last visit

**Navigation**
- Visits duration, File type, Viewed, Full list, Entry, Exit
- Operating Systems, Versions, Unknown
- Browsers, Versions, Unknown
- Referers:
  - Origin, Referring search engines, Referring sites
  - Search, Search Keyphrases, Search Keywords

**Last Update:** 14 Jan 2008 - 02:59
**Reported period:** [Year] 2007 OK

**Back to main page**

### Robots/Spiders visitors

<table>
<thead>
<tr>
<th>30 different robots</th>
<th>Hits</th>
<th>Bandwidth</th>
<th>Last visit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Googlebot</td>
<td>1393868+104</td>
<td>5.11 GB</td>
<td>31 Dec 2007 - 23:50</td>
</tr>
<tr>
<td>Inktomi Slurp</td>
<td>36668+221</td>
<td>554.25 MB</td>
<td>31 Dec 2007 - 23:55</td>
</tr>
<tr>
<td>MSNBot</td>
<td>19522+2</td>
<td>699.90 MB</td>
<td>28 Dec 2007 - 08:01</td>
</tr>
<tr>
<td>Unknown robot</td>
<td>15949+13</td>
<td>89.34 MB</td>
<td>31 Dec 2007 - 22:24</td>
</tr>
<tr>
<td>AskJeeves</td>
<td>7016+1</td>
<td>106.29 MB</td>
<td>31 Dec 2007 - 23:49</td>
</tr>
<tr>
<td>Google Adsense</td>
<td>2701</td>
<td>100.26 MB</td>
<td>31 Dec 2007 - 22:10</td>
</tr>
<tr>
<td>psbot</td>
<td>2268+1</td>
<td>80.48 MB</td>
<td>31 Dec 2007 - 09:59</td>
</tr>
<tr>
<td>Unknown robot</td>
<td>930+1</td>
<td>19.10 MB</td>
<td>31 Dec 2007 - 09:34</td>
</tr>
<tr>
<td>Turn It In</td>
<td>350+1</td>
<td>6.32 MB</td>
<td>03 Sep 2007 - 15:44</td>
</tr>
<tr>
<td>BaiduSpider</td>
<td>300</td>
<td>10.22 MB</td>
<td>26 Nov 2007 - 07:32</td>
</tr>
<tr>
<td>GigaBot</td>
<td>243</td>
<td>5.27 MB</td>
<td>30 Dec 2007 - 05:06</td>
</tr>
<tr>
<td>Scooter</td>
<td>90+3</td>
<td>288.75 KB</td>
<td>27 Nov 2007 - 14:30</td>
</tr>
<tr>
<td>PhpDig</td>
<td>91</td>
<td>2.28 MB</td>
<td>21 Oct 2007 - 09:51</td>
</tr>
<tr>
<td>WISENutbot</td>
<td>76</td>
<td>1.94 MB</td>
<td>13 Jan 2007 - 14:04</td>
</tr>
<tr>
<td>Magpie</td>
<td>25</td>
<td>43.48 KB</td>
<td>24 Dec 2007 - 00:51</td>
</tr>
<tr>
<td>Unknown robot</td>
<td>0+16</td>
<td>4.38 KB</td>
<td>14 Nov 2007 - 03:43</td>
</tr>
<tr>
<td>Echo!</td>
<td>14</td>
<td>287.09 KB</td>
<td>27 Dec 2007 - 13:56</td>
</tr>
<tr>
<td>Internet Shinchakubin</td>
<td>13</td>
<td>385.03 KB</td>
<td>27 Nov 2007 - 15:23</td>
</tr>
<tr>
<td>BBot</td>
<td>10</td>
<td>146.35 KB</td>
<td>13 Jun 2007 - 15:17</td>
</tr>
<tr>
<td>arks</td>
<td>8</td>
<td>142.24 KB</td>
<td>27 Nov 2007 - 12:25</td>
</tr>
<tr>
<td>MSIECrawler</td>
<td>8</td>
<td>263.02 KB</td>
<td>26 Dec 2007 - 11:16</td>
</tr>
<tr>
<td>The Python Robot</td>
<td>5</td>
<td>429.04 KB</td>
<td>23 Nov 2007 - 08:34</td>
</tr>
</tbody>
</table>
**Politeness?**

### Monthly history

<table>
<thead>
<tr>
<th>Month</th>
<th>Unique visitors</th>
<th>Number of visits</th>
<th>Pages</th>
<th>Hits</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 2007</td>
<td>1221</td>
<td>2946</td>
<td>8938</td>
<td>30536</td>
<td>699.28 MB</td>
</tr>
<tr>
<td>Feb 2007</td>
<td>1179</td>
<td>3099</td>
<td>7852</td>
<td>20475</td>
<td>415.75 MB</td>
</tr>
<tr>
<td>Mar 2007</td>
<td>1120</td>
<td>3063</td>
<td>7099</td>
<td>18978</td>
<td>350.88 MB</td>
</tr>
<tr>
<td>Apr 2007</td>
<td>1362</td>
<td>3067</td>
<td>7175</td>
<td>30320</td>
<td>599.91 MB</td>
</tr>
<tr>
<td>May 2007</td>
<td>1612</td>
<td>3746</td>
<td>7584</td>
<td>25114</td>
<td>469.32 MB</td>
</tr>
<tr>
<td>Jun 2007</td>
<td>1474</td>
<td>3662</td>
<td>7138</td>
<td>22292</td>
<td>370.11 MB</td>
</tr>
<tr>
<td>Jul 2007</td>
<td>1592</td>
<td>4210</td>
<td>9165</td>
<td>24766</td>
<td>430.61 MB</td>
</tr>
<tr>
<td>Aug 2007</td>
<td>1658</td>
<td>4567</td>
<td>10600</td>
<td>24142</td>
<td>336.08 MB</td>
</tr>
<tr>
<td>Sep 2007</td>
<td>1458</td>
<td>4403</td>
<td>11149</td>
<td>24414</td>
<td>356.60 MB</td>
</tr>
<tr>
<td>Oct 2007</td>
<td>2148</td>
<td>5299</td>
<td>12877</td>
<td>36427</td>
<td>783.78 MB</td>
</tr>
<tr>
<td>Nov 2007</td>
<td>2890</td>
<td>6317</td>
<td>15300</td>
<td>40487</td>
<td>833.75 MB</td>
</tr>
<tr>
<td>Dec 2007</td>
<td>2748</td>
<td>6631</td>
<td>17553</td>
<td>42281</td>
<td>1.03 GB</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>20462</strong></td>
<td><strong>51010</strong></td>
<td><strong>122430</strong></td>
<td><strong>340232</strong></td>
<td><strong>6.55 GB</strong></td>
</tr>
</tbody>
</table>
### Politeness?

#### Last Update:
20 Jan 2009 - 03:14

#### Reported period:
- Year - 2008

---

#### Robots/Spiders visitors

<table>
<thead>
<tr>
<th>Robot/Spiders</th>
<th>Hits</th>
<th>Bandwidth</th>
<th>Last visit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Googlebot</td>
<td>110194+158</td>
<td>1.95 GB</td>
<td>31 Dec 2008 - 23:43</td>
</tr>
<tr>
<td>Inktomi Slurp</td>
<td>13689+156</td>
<td>244.14 MB</td>
<td>31 Dec 2008 - 23:38</td>
</tr>
<tr>
<td>AskJeeves</td>
<td>6344+3</td>
<td>113.74 MB</td>
<td>30 Dec 2008 - 02:55</td>
</tr>
<tr>
<td>MSNBot</td>
<td>5014+11</td>
<td>146.12 MB</td>
<td>31 Dec 2008 - 22:36</td>
</tr>
<tr>
<td>Unknown robot (identified by 'crawl')</td>
<td>4302+5</td>
<td>47.72 MB</td>
<td>31 Dec 2008 - 03:04</td>
</tr>
<tr>
<td>Google AdSense</td>
<td>3005</td>
<td>112.90 MB</td>
<td>31 Dec 2008 - 21:14</td>
</tr>
<tr>
<td>Unknown robot (identified by 'robot')</td>
<td>2518+138</td>
<td>89.74 MB</td>
<td>31 Dec 2008 - 15:15</td>
</tr>
<tr>
<td>psbot</td>
<td>1263</td>
<td>27.87 MB</td>
<td>17 Dec 2008 - 07:56</td>
</tr>
<tr>
<td>GigaBot</td>
<td>261</td>
<td>7.14 MB</td>
<td>30 Dec 2008 - 18:05</td>
</tr>
<tr>
<td>Unknown robot (identified by 'spider')</td>
<td>201</td>
<td>9.23 MB</td>
<td>29 Dec 2008 - 03:30</td>
</tr>
<tr>
<td>Turn It In</td>
<td>178+1</td>
<td>3.34 MB</td>
<td>24 Nov 2008 - 13:57</td>
</tr>
<tr>
<td>BBot</td>
<td>84</td>
<td>1.65 MB</td>
<td>24 Nov 2008 - 19:42</td>
</tr>
<tr>
<td>The Python Robot</td>
<td>74+3</td>
<td>3.12 MB</td>
<td>31 Aug 2008 - 11:58</td>
</tr>
<tr>
<td>Unknown robot (identified by hit on 'robots.txt')</td>
<td>0+54</td>
<td>15.04 KB</td>
<td>28 Dec 2008 - 19:38</td>
</tr>
<tr>
<td>EchOl</td>
<td>48</td>
<td>1.14 MB</td>
<td>28 Nov 2008 - 13:02</td>
</tr>
<tr>
<td>Magpie</td>
<td>43</td>
<td>21.74 KB</td>
<td>17 Nov 2008 - 11:16</td>
</tr>
<tr>
<td>MSIECrawler</td>
<td>42</td>
<td>561.59 KB</td>
<td>01 Dec 2008 - 23:41</td>
</tr>
<tr>
<td>arks</td>
<td>39</td>
<td>438.47 KB</td>
<td>21 Dec 2008 - 04:30</td>
</tr>
<tr>
<td>SurveyBot</td>
<td>21</td>
<td>14.07 KB</td>
<td>29 Dec 2008 - 23:18</td>
</tr>
<tr>
<td>Internet Shinchakubin</td>
<td>17</td>
<td>267.64 KB</td>
<td>26 Aug 2008 - 06:55</td>
</tr>
<tr>
<td>Scooter</td>
<td>8</td>
<td>16 Bytes</td>
<td>24 Apr 2008 - 00:14</td>
</tr>
<tr>
<td>BaiduSpider</td>
<td>7</td>
<td>204.01 KB</td>
<td>25 Apr 2008 - 19:46</td>
</tr>
<tr>
<td>Fish search</td>
<td>6</td>
<td>34.53 KB</td>
<td>08 Apr 2008 - 15:46</td>
</tr>
<tr>
<td>WISEN-arkbot</td>
<td>4</td>
<td>26.51 KB</td>
<td>18 May 2008 - 23:54</td>
</tr>
</tbody>
</table>

---

#### Data history

- **Number of visits**
- **Pages**
- **Hits**
- **Bandwidth**

<table>
<thead>
<tr>
<th>Date</th>
<th>Number of visits</th>
<th>Pages</th>
<th>Hits</th>
<th>Bandwidth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apr 2008</td>
<td>7047</td>
<td>17437</td>
<td>40911</td>
<td>1.03 GB</td>
</tr>
<tr>
<td>May 2008</td>
<td>8477</td>
<td>18878</td>
<td>47929</td>
<td>1.08 GB</td>
</tr>
<tr>
<td>Jun 2008</td>
<td>9001</td>
<td>22892</td>
<td>57030</td>
<td>1.32 GB</td>
</tr>
<tr>
<td>Jul 2008</td>
<td>11510</td>
<td>26871</td>
<td>61373</td>
<td>1.29 GB</td>
</tr>
<tr>
<td>Aug 2008</td>
<td>12405</td>
<td>31227</td>
<td>57028</td>
<td>947.39 MB</td>
</tr>
<tr>
<td>Sep 2008</td>
<td>8088</td>
<td>25365</td>
<td>53173</td>
<td>1.08 GB</td>
</tr>
<tr>
<td>Oct 2008</td>
<td>7275</td>
<td>20348</td>
<td>79872</td>
<td>1.56 GB</td>
</tr>
<tr>
<td>Nov 2008</td>
<td>6212</td>
<td>24236</td>
<td>55827</td>
<td>2.11 GB</td>
</tr>
<tr>
<td>Dec 2008</td>
<td>6593</td>
<td>40161</td>
<td>76103</td>
<td>3.00 GB</td>
</tr>
<tr>
<td>Jan 2009</td>
<td>10887</td>
<td>45152</td>
<td>110285</td>
<td>3.19 GB</td>
</tr>
<tr>
<td>Feb 2009</td>
<td>12143</td>
<td>29037</td>
<td>65801</td>
<td>1.36 GB</td>
</tr>
<tr>
<td>Mar 2009</td>
<td>6724</td>
<td>17717</td>
<td>47438</td>
<td>924.81 MB</td>
</tr>
<tr>
<td>Apr 2009</td>
<td>106362</td>
<td>319320</td>
<td>752050</td>
<td>18.84 GB</td>
</tr>
</tbody>
</table>
Robots.txt - Exclusion

- Protocol for giving spiders ("robots") limited access to a website
  - Source: [http://www.robotstxt.org/wc/norobots.html](http://www.robotstxt.org/wc/norobots.html)
  - Website announces what is okay and not okay to crawl:
    - This file holds the restrictions
Robots.txt Example


```plaintext
User-agent: MOMspider
Disallow: /cgi-bin/
Disallow: /Admin/MOM/
Disallow: ~/fielding/MOM/
Disallow: /TR/
Disallow: /Server/
Disallow: /Document/
Disallow: /MetaServer/
Disallow: ~/eppstein/pubs/cites/
Disallow: ~/fiorello/pvt/

User-agent: *
Disallow: /cgi-bin/
Disallow: /Test/
Disallow: /Admin/
Disallow: /TR/
Disallow: /Server/
Disallow: /Document/
Disallow: /MetaServer/
Disallow: ~/fielding/MOM/
Disallow: ~/kanderso/hidden
Disallow: ~/eppstein/pubs/cites/
Disallow: ~/fiorello/pvt/
Disallow: ~/dean/
Disallow: ~/wwwoffice/
Disallow: ~/ucounsel/
Disallow: ~/sao/
Disallow: ~/support/
Disallow: ~/icsdb/
Disallow: ~/bin/
```

• The Multi-Owner Maintenance Spider
• Script files
• Local MOMspider output
• Local MOMspider output
• Dienst Technical Report Server
• Dienst Technical Report Server
• Dienst Technical Report Server
• Dienst Technical Report Server
• Dienst Technical Report Server
• Dienst Technical Report Server
• Dienst Technical Report Server
• Eppstein Database
• Private pages

• All other spiders should avoid
• Script files
• The test area for web experimentation
• Huge server statistic logs
• Dienst Technical Report Server
• Dienst Technical Report Server
• Dienst Technical Report Server
• Dienst Technical Report Server
• Dienst Technical Report Server
• Local MOMspider output
• Ken Anderson's stuff
• Eppstein Database
• Private pages
Sitemaps - Inclusion


```xml
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
  <url>
    <loc>http://www.example.com/</loc>
    <lastmod>2005-01-01</lastmod>
    <changefreq>monthly</changefreq>
    <priority>0.8</priority>
  </url>
  <url>
    <loc>http://www.example.com/catalog?item=12&amp;desc=vacation_hawaii</loc>
    <changefreq>weekly</changefreq>
  </url>
  <url>
    <loc>http://www.example.com/catalog?item=73&amp;desc=vacation_new_zealand</loc>
    <lastmod>2004-12-23</lastmod>
    <changefreq>weekly</changefreq>
  </url>
  <url>
    <loc>http://www.example.com/catalog?item=74&amp;desc=vacation_newfoundland</loc>
    <lastmod>2004-12-23T18:00:15+00:00</lastmod>
    <priority>0.3</priority>
  </url>
  <url>
    <loc>http://www.example.com/catalog?item=83&amp;desc=vacation_usa</loc>
    <lastmod>2004-11-23</lastmod>
  </url>
</urlset>
```